DOCUMENT RESUME

ED 393 888 TM 024 805

AUTHOR Borman, Geoffrey D.; D'Agostino, Jerome V.
TITLE Title I and Student Achievement: A Meta-Analysis of 30 Years of Test Results.
PUB DATE Apr 95
NOTE 39p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 18-22, 1995).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Academic Achievement; Achievement Tests; Comparative Analysis; *Compensatory Education; Educational Finance; Educational Improvement; Educationally Disadvantaged; Elementary Secondary Education; Financial Support; High Risk Students; *Meta Analysis; *Program Evaluation; School Districts; State Departments of Education; Supplementary Education; *Test Results
IDENTIFIERS *Elementary Secondary Education Act Title I

ABSTRACT

Title I of the Elementary and Secondary Education Act was implemented in 1965 to provide financial assistance to state and local education agencies to meet the special needs of educationally disadvantaged children. Federal funding supports a variety of supplemental services that share the collective purpose of improving educational opportunities and outcomes for low-achieving students from schools with concentrations of poverty. Evaluation of Title I programs has usually been through a norm-referenced model based on achievement test result changes. This meta-analysis, which included 17 studies, considered whether program services, taken as a whole, had significant impact on student achievement; and it offered the possibility of determining whether choices of different testing cycles and comparison methods have provided differential estimates of the program's impact. Evidence from the analysis indicated that Title I has not fulfilled its original expectation of closing the achievement gap between at-risk students and their more advantaged peers. Results did suggest that without the program it is likely that children served over the last 30 years would have fallen further behind academically. Evaluated from this perspective, Title I has been an invaluable supplement to schools serving disadvantaged children. (Contains 4 tables, 10 figures, and 50 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 393 888

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
JEROME V. D'AGOSTINO

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)"

Title I and Student Achievement:

A Meta-Analysis of 30 Years of Test Results

Geoffrey D. Borman **Jerome V. D'Agostino**

University of Chicago

Paper presented at the American Educational Research Association Annual Meeting, San Francisco, CA, April 1995.

TM 02 4 805

BEST COPY AVAILABLE

The authors offer sincere thanks to the following personnel from the U.S. Department of Education, who were instrumental in locating key documents and in providing helpful advice in our search of the Title I evaluation literature:

Judith Anderson
Joanne Bogart
Alan Ginsburg
Ed Glassman
Val Plisko
Elois Scott
Bob Stonehill

In addition, we thank the following Title I researchers who provided other reports and needed information:

Jim Fortune
Ronald Hambleton
Mary Kennedy
Donald McLaughlin
Sol Pelavin
Judy Pfannestiel
Beth Sinclair

Finally, we thank Larry Hedges for his technical and statistical advice.

Title I of the Elementary and Secondary Education Act (ESEA) was implemented in 1965 "to provide financial assistance to state and local education agencies to meet the special needs of educationally deprived children" [Section 552, Education Consolidation and Improvement Act (ECIA), 1981]. This federal funding source has supported a variety of supplemental services that share a collective purpose: to improve educational opportunities and outcomes for low-achieving students from schools with concentrations of poverty. Since the inception of Title I, states and districts have collected annual evaluative and descriptive information regarding the local operation of their programs. At the federal level, researchers have attempted to produce national estimates of participants' achievement gains by synthesizing the results of these state and district test data. In addition, the federal government has sponsored two systematic, nationally-representative, longitudinal assessments of participants' achievement, the *Sustaining Effects Study (SES)* and the ongoing *Prospects* study, and a number of smaller studies, which have examined the implementation and effectiveness of the program from a variety of perspectives. Despite this seeming wealth of information, the overall educational effectiveness of the program has remained as a matter of debate. In reviewing the evolution of Title I evaluation research, many of the factors that have contributed to this uncertainty become apparent.

Title I was the first federal education law to mandate annual effectiveness evaluations (Timpane, 1976). However, with the political urgency to respond to the rising demand for the improvement of human rights and for the increased development of underprivileged populations, there was little time for planning and for developing specific implementation and evaluation guidelines and regulations (Gordon, 1970). In 1968, three years after the inception of Title I, the federal government initiated the *Belmont* evaluative information system, which was designed to provide annual reports of cost, process, pupil background, and achievement data (Timpane, 1976). The production of these reports involved each local education agency (LEA) sending its locally developed evaluation to its state education agency (SEA). The states, in turn, attempted to produce summaries of the various

district evaluations and submitted them to the U.S. Office of Education where the general findings were compiled as a final document for Congress.

However, given the differences in the level of technical sophistication among researchers at the local level, and given the infancy of the federal program evaluation model, the quality and methodological approaches of the district reports varied dramatically (Wargo, Tallmadge, Michaels, Lipe, and Morris, 1972). Although the receipt of Title I funds was conditional upon LEAs providing annual assessments of their Title I students' educational achievement in basic skills, during this early period of the program, the data were far from uniform. Attempts to aggregate and synthesize the achievement results at the national level were largely unsuccessful. In many other cases, districts were uncooperative, and did not respond to requests from the states for achievement test results. Indeed, the few early federal attempts to gauge the educational effectiveness of the program on a national scale were complete failures due to the lack of sufficient local achievement data. For instance, the attempt by Glass, *et al.* (1970) to collect test scores from existing school records revealed that only 7,784 or (7.5 percent) matched individual pre- and posttest records were available for the 104,080 students in the 1968-69 national survey.

Instead of estimating the impact of the program on student achievement, the value of the early federal evaluations was in addressing a more fundamental question: Were the federal funds being spent on the targeted students for the intended purpose of providing some form of supplemental educational services? The Washington Research Project was one of the more prominent studies that revealed large-scale violations in the operation of the program (Martin and McClure, 1970). Similarly, Wargo, *et al.* (1972) concluded that localities had disregarded regulations, guidelines, and program criteria and had not implemented Title I as intended by Congress. Farrar and Millsap (1986) estimated that during the first five years of the program, states and districts misappropriated as much as 15 percent of the federal dollars. These revelations prompted the development of more stringent federal

guidelines and program compliance measures. During the 1970s, as program administrators took on more active monitoring roles that were necessary to oversee the distribution of millions of federal dollars each year, localities more faithfully implemented Title I. Blatant misappropriations of funds were no longer a key concern of Title I policymakers and stakeholders. As a result, research priorities began to evolve. When Title I was reauthorized in 1974 under P.L. 93-380, Congress took the initiative to develop systematic means for evaluating the educational effectiveness of the program on a national basis. This law included requirements specifying that objective criteria should be utilized in the evaluation of all programs, and that methodological approaches should be developed that yield comparable data on a state-wide and nation-wide basis.

Under federal contract, RMC Research Corporation developed three evaluation models that fulfilled these requirements (Tallmadge and Wood, 1981). These models included a norm-referenced comparison, a control group comparison, and a regression-discontinuity design. Although some school districts utilized any one of these three evaluation methods during the late 1960s and the 1970s, it was not until the 1979-80 school year that the SEAs were required to compile results obtained from these models and submit them to the U.S. Department of Education. This intergovernmental system, known as the Title I Evaluation and Reporting System (TIERS), has permitted a systematic compilation of the results of various federally-approved standardized tests from 1979-80 to the present. Despite the change from a mandatory to a voluntary reporting system with the approval of the Education Consolidation and Improvement Act (ECIA) of 1981, states have continued to submit their Title I student achievement data to the U.S. Department of Education on a somewhat reliable basis.

Because the methodological and analytical procedures of the control group and regression discontinuity models were more complicated, the norm-referenced model became the evaluation method of choice at the district level. According to this model, Title I programs have been evaluated based on pre-post change scores from various standardized achievement tests administered on either a

fall-to-spring or annual testing cycle. If the mean change score of participating students within a school is greater than 0 normal curve equivalents (NCEs), which are normalized percentile scores with a mean of 50 and a standard deviation of 21.06, the program is said to be effective. A mean gain greater than 0 NCEs has been interpreted as evidence of programmatic impact, on the assumption that in the absence of Title I instruction students tend to remain at the same national percentile rank over time -- the "equipercentile assumption" (Tallmadge and Wood, 1981).

The validity of this model has been questioned by a number of researchers. For example, Linn (1979) found a lack of support for the equipercentile assumption and concluded that gains may be inflated by regression effects, especially given Title I students' typically low pretest scores. On the other hand, some empirical results have provided support for the assumption. Although the researchers recognized that the regression problem remained, they reached similar conclusions regarding its effect on the national results. Kaskowitz and Norwood (1977) found deviations from equal percentile estimates ranging from -0.2 to 2.9 in reading and from -1.2 to 5.5 in mathematics on national data from the Metropolitan Achievement Test. For minority students whose pretest scores were moderately low, they found that the equipercentile assumption led to a slight overestimate of no-treatment growth. Likewise, Tallmadge (1982), using nationally-representative longitudinal data from the *Sustaining Effects Study* (*SES*) and from the California Achievement Test norming sample, found that the mean deviation of longitudinal gains from equal percentile estimates was very close to 1 NCE in both samples. Similar to Tallmadge's (1982) assessments, Linn (1980) estimated the impact of regression on nationally-aggregated gains to be about 2 NCEs at second grade, and 1 NCE for third grade and above.

In addition to potential regression effects, Linn, Dunbar, Harnisch, and Hastings (1982) noted that in some cases teachers and administrators may produce erroneous conversions of raw scores to NCEs and may vary pre- and posttest conditions in subtle ways in hopes of inflating the gains posted

by the students in their programs. However, the authors mentioned that manual conversions to NCEs based on table look-ups were the main cause of conversion errors. With the widespread availability of computers in more recent years, this source of error has surely decreased. Regarding the potential biases introduced by varying administration conditions, Linn, *et al.*(1982) suspected that flagrant practices, such as altering testing time limits, were extremely rare. Instead, encouraging pep talks by teachers preceding the administration of posttests may be the most frequent infraction. As with most local testing programs, practice effects may introduce another source of positive bias in student gains that reflect improved test taking skills or successful teaching to the test (Linn, 1982). Although these problems may have some influence on the national data, their extent and impact is not known.

Other factors may affect the stability and precision of the NCE metric, but do not appear to produce systematic biases of the national results. First, Jaeger (1979) reported that NCE gains based on the various nationally-normed tests administered to Title I participants were not equivalent on a consistent basis. Although the NCE provides a common metric, the data reported by districts are susceptible to small differences attributable to variability in standardized tests' norms and actual content. Second, as mentioned earlier, the overall quality of data submitted by some states has been inadequate, and a few states have declined to participate in the TIERS system since it became voluntary with the approval of the ECIA legislation in 1981. Nevertheless, Kennedy, *et al.* (1986) claimed the achievement patterns reported by TIERS are consistent from year to year suggesting that problems related to data quality have not systematically biased achievement upward or downward.

Aside from these technical issues, many researchers have demonstrated that different testing cycles may lead to different estimates of the program's effectiveness. In large part, these differing results seem to be attributable to the "summer effect" (David and Pelavin, 1977; Heyns, 1977; Thomas and Pelavin, 1976; Hayes and Grether, 1969). As David and Pelavin (1977) have demonstrated, large achievement gains by Title I students over the school year typically are followed

by diminished summer growth, or by achievement losses. These findings suggest that although standardization procedures of achievement tests assume a slower rate of growth during the summer months, disadvantaged students typically achieve no gain over the summer or may even lose ground when compared to these normed gain standards. Therefore, due to the impact of the intervening summer months, it is reasonable to expect that Title I students will tend to post smaller annual gains than fall-to-spring gains.

Slavin, Karweit, and Madden (1989), on the other hand, claimed that when Title I students' achievement growth is measured on a fall-to-spring testing cycle it is greatly overestimated by an "unknown artifact." The authors stated the most obvious indication of this is that national gains have averaged around 7 NCEs when measured fall-to-spring, but only about 3 NCEs when measured on a spring-to-spring, or annual, cycle. The positive biases resulting from potential differences in the administration of pre- and posttests, which were mentioned above, may be part of the problem to which the authors refer. When local evaluations are based on an annual cycle, they typically employ the posttest as the pretest for the following school year. Therefore, with an annual testing cycle, the incentive for local administrators and teachers to alter testing schedules and conditions is reduced. Although the role of the summer effect has been more clearly established, some of the potential artifacts discussed above may contribute to the differences between annual and fall-to-spring gains.

The control group model proposed by RMC Research probably provides the most accurate estimates of program effectiveness, but it has not been implemented frequently. With the rather widespread availability of state and federal compensatory education services for at-risk students, the difficulty in identifying comparable control students often has plagued attempts to measure the impact of Title I (Slavin, *et al.*, 1989). Random assignment to the program would not be legal nor would it be ethical, therefore, evaluations using controls have employed quasi-experimental methods with differing control group definitions and criteria.

The goal of "narrowing the achievement gap" between disadvantaged children and their more advantaged peers has two distinct definitions that have influenced the selection of an appropriate control group. Some researches have responded to the question: Does participation in Title I narrow the achievement difference between program participants and a nationally-representative sample of non-participants? Alternatively, others have questioned whether this gap would widen without the existence of Title I services. Researchers have responded to the former question by comparing Title I students to all other non-participating students, while the latter question has been addressed by comparing Title I students to similarly disadvantaged controls who did not receive compensatory education services. However, these types of comparisons are very rare at the district level and have been implemented in only two nationally-representative assessments of the program: the *SES* and the ongoing *Prospects* evaluation.

Previous Research Findings

Much of the conventional wisdom concerning the educational effectiveness of Title I has been derived from federally-sponsored studies and from the compilation of state reports under the TIERS model. To date, findings from the nationally-representative *SES* have contributed the most to these widely held beliefs. The study, conducted from 1976 to 1979, reported the following central finding:

> Participating students outperformed similarly disadvantaged students who did not receive program services, but they did not attain the levels of academic achievement of their more advantaged peers (Carter, 1981).

Some analysts have interpreted this result to indicate that the program has had a modest effect on student achievement, but has not fulfilled its original intention to raise the achievement of its participants to the level of their more economically and educationally advantaged peers. Other researchers have suggested that without Title I services, its beneficiaries would be far worse off, and that it is naive to expect a single supplemental educational program to compensate for deficiencies in the at-risk student's total educational environment.

Also based on data collected for the *SES*, Carter (1981) concluded that Title I students achieved greater gains in math programs than in reading, and greater gains in the earlier grades than in the upper grades. These results seem to be corroborated by the TIERS performance summaries collected from 1979 to the present. Given that approximately 90 percent of Title I participants are enrolled in grades K through 8 (Kennedy, *et al.*, 1986), and that most districts allocate the majority of their program funds to these grades, it is not surprising that children in the earlier grades have made greater gains. Yet this explanation does not account for the apparent success of math programs, because more Title I funds have been spent on reading than math. Despite the large differences in student outcomes between annual and fall-spring testing schedules, Kennedy, *et al.* (1986) pointed out that these achievement patterns, by subject and by grade, are consistent across both cycles.

Previous evidence regarding the historical variation of the Title I program's effects on student achievement is limited. The results from the National Assessment of Educational Progress (NAEP) have suggested that the achievement gaps between both African American and white students and economically disadvantaged and advantaged students diminished during the 1980s. Furthermore, students from schools that received federal Title I funds made significantly greater gains between 1970 and 1980 than pupils from non-Title I eligible schools in all three grades tested by NAEP (Forbes, 1985). Although compensatory education probably played some role in this apparent improvement, the data are aggregated at the school-level and it is not possible to compare the test scores of Title I students to non-participants. The NAEP results may indicate a positive trend for the effect of Title I, yet Davis (1991) indicated that the student gains reported under the TIERS model have been consistent in magnitude from year to year. No studies have attempted to collect the data necessary to investigate the trends in participant achievement data across the life of the program.

There have been six known attempts over the years to synthesize the results of Title I student achievement evaluations (Wargo, *et al.*, 1972; McLaughlin, 1977; Rossi, *et al.*, 1977; Mullin and

Summers, 1983; Kennedy, *et al.*, (1986); Slavin, *et al.*, 1989). The review by Kennedy, *et al.*, (1986), which was the Second Interim Report from the 1986 National Assessment of Chapter 1, provided an excellent overview of the evaluation results from the first twenty years of the program. Several of these findings were discussed above. However, the U.S. Department of Education has sponsored considerable research since the completion of the report in 1986. The review by Slavin, *et al.* (1989) was more contemporary, and utilized "best-evidence synthesis" techniques and effect sizes. Yet, the authors made no attempt to provide an overall statistical appraisal of the effectiveness of Title I based on data representative of typical programs. Instead, the researchers documented the effects and qualitative characteristics of a variety of "exemplary" compensatory education programs.

Like the Kennedy, *et al.*, (1986) report, the earlier studies are better defined as reviews of the existing evaluation literature than as quantitative meta-analyses. All but one review, Mullin and Summers (1983), pre-dated the TIERS model and compiled data from central federal studies only. Most importantly, the Wargo, *et al.* (1972), McLaughlin (1975), and Rossi, *et al.* (1977) studies documented the failure of federal evaluations to gather sufficient test score data that were nationally-representative, and the lack of methodological rigor in estimating program effectiveness. Overall, the conclusions of these early reviews implied that the program had not been implemented effectively and had little impact on closing the achievement gap.

The more recent study by Mullin and Summers (1983) offered the most comprehensive analysis of the effectiveness of compensatory education. However, this synthesis concentrated on the specific relationship between amounts of expenditures and levels of student achievement, and was not specific to Title I but included a variety of state and federal compensatory programs. After extensive review of the studies that they accumulated, the authors concluded that compensatory education participants had a slight edge over non-participants, but that the effects were not sustained, and that there was no relationship between dollars spent and achievement gains.

BEST COPY AVAILABLE

In contrast to previous reviews of the Title I evaluation literature, we employed quantitative meta-analytic techniques to produce an overall assessment of the impact of the program on student achievement. This approach permitted a systematic interpretation of the moderating effects of different evaluation methods and program characteristics on Title I effect size estimates. The data were derived from TIERS (academic years 1979-80 through 1992-93) and from central federal studies, which ranged from the second year of the program, 1966, to 1992-93. The main purposes of this analysis are:

1. to quantify the overall impact of Title I;
2. to evaluate variation in program effect sizes over time;
3. to examine how different control group definitions affect estimates of program impact;
4. to evaluate the variation in achievement associated with different testing cycles (i.e., fall-to-spring versus annual);
5. to compare the effectiveness of math programs to the effectiveness of reading programs;
6. to assess differences in programmatic effects across grades.

It is predicted that the conventional wisdom, which suggests a modest overall program impact, will be confirmed. Other findings from *SES* and TIERS pertaining to achievement differences across reading and mathematics programs, as well as across grade levels, are also predicted to be substantiated. The methodological issues of control group definition and testing cycle are hypothesized to affect estimations of program impact. Comparisons of Title I students' achievement results to norms and to similarly needy students' outcomes are anticipated to produce more favorable estimates than comparisons to non-participants' achievement results. Also, primarily due to the summer effect, fall-to-spring testing cycles are expected to provide higher estimates of positive program effects when compared to the results from annual cycles. Finally, although no systematic analysis has addressed the potential historical variation in the effect of Title I on student achievement, it is hypothesized that the program has had a relatively stable effect over time.

Methods

Literature Search Methods

Broad searches of the literature on compensatory education and its effects on student achievement were conducted. Methods included preliminary computerized searches of the ERIC database (1966-1995), a second phase of exhaustive "footnote chasing" based on all reports found through the ERIC system, and a third phase of numerous consultations with current and former Title I researchers and with personnel from the U.S. Department of Education. In scanning the ERIC database, free-text searches were conducted using the general key words: "compensatory education" and "achievement." In subsequent searches of ERIC, the following key words were used alone, and in various combinations: "Chapter 1," Title I," "Elementary and Secondary Education Act," and "Educational Consolidation and Improvement Act." These terms were employed in various combinations with other key words: "federal," "evaluation," "outcomes," "gains," "longitudinal," "effectiveness," "bibliography," "review," "synthesis," and "meta-analysis." The general literature search also relied on the examination of reference lists from all studies initially found. Finally, personnel at the U.S. Department of Education and various researchers from across the country were instrumental in providing additional "fugitive" reports that were not available through conventional database searches. After scrutinizing the reference lists of all studies, no other available evaluations of Title I and student achievement outcomes were found.

Inclusion Criteria

Due to the Title I three-tiered evaluation reporting system, the vast majority of research conducted since 1979 can be reviewed based on three distinct sources: district, state, or national summary reports. The federal summaries of local evaluations conducted under TIERS were selected as the primary source for three reasons. First, federal summaries were more easily accessible than the thousands of district and state reports that have been conducted over the years.. Second, inclusion of a

mixture of reports from all three levels within a given year would result in the synthesis of duplicative data. Third, at the federal level, personnel from the U.S. Department of Education, and/or their contractors, perform an edit check review of the state reports and attempt to reconcile anomalous results. Prior to the advent of TIERS, all federally-commissioned evaluations that attempted to summarize the effects of the program were considered for inclusion. The selection of the federal level as the unit of analysis most directly supported the attempt to produce an overall estimate of the effect of Title I programs on student achievement.

Liberal inclusion criteria were applied in the preliminary stages of the literature search. All study abstracts provided by the ERIC database searches and all evaluations of Title I student achievement that were referenced in the ERIC documents were read to ascertain whether any report of Title I test score data may be provided by the studies. If a study did not report these data, it was excluded from further consideration. Over 150 studies, abstracts, and summaries were read during the preliminary stages of the review process, and 81 were selected for the second stage of review.

In the second stage, studies were deemed admissible based on the following eligibility criteria:

- number of students tested was provided;
- sufficient test score data for Title I and, when applicable, comparison group students were provided from which effect sizes could be computed;
- matched pre- and posttest scores, or individual gain scores were available;
- the treatment group received federal Title I services;
- test score data were reported separately for reading and/or mathematics;
- data were not duplicated in another study accepted for inclusion;
- test score data were reported separately by grade level;
- data were from representative, non-exemplary programs;
- the researchers attempted to collect data that would provide the basis for national estimates of the program's effectiveness and/or the evaluation was commissioned by the federal government.

Most studies reviewed in the second phase did not meet these eligibility requirements. In large part, this was because studies were based on the same data. Due to the choice of the federal level as the unit of analysis, numerous district and state evaluations were not considered for the meta-analysis

because they were included in the national TIERS state performance summaries. Some of the federal evaluations were not selected for other reasons. First, many were reanalyses of data from evaluations that were already selected for the synthesis. For instance, there were a number of studies that reanalyzed the *SES*, however, the Kennedy, *et al.*, (1986) report was selected because it provided test scores in the NCE metric. Second, a number of studies, such as the *It Works!* series, sampled students from exemplary Title I programs only. Third, other federal evaluations did not distinguish Title I recipients from students receiving other forms of compensatory education (e.g., the *Reading Study* conducted by the Educational Testing Service). Of the 81 studies to which these criteria were applied, 17 met all requirements and were selected for analysis.

Characteristics of Selected Studies

Only Frechtling's (1978) report, based on data from the *Instructional Dimensions Study*, provided test scores by Title I program model. No other studies collected information that would permit comparisons among possible program variations, such as pullout and in-class instruction. Two program distinctions that *were* possible included instructional subject area, reading and math, and grade level, 1 to 12. Because all studies included separate reports of reading and math test scores by grade, it is assumed that students received some Title I instruction in the subject for which test scores were reported.

In all 17 studies, standardized tests were administered to measure student achievement outcomes. The evaluations either reported the actual test that was used, or summarized data across various standardized tests based on a common metric -- NCEs or grade equivalents. Students were tested on either fall-to-spring or annual cycles. Thus, all studies examined the effectiveness of the program with a two-wave, pretest-posttest design. Although some researchers (e.g., David and Pelavin, 1977) have considered a gain achieved during a single annual testing cycle as a type of

"sustained effect," an estimate of the program's longitudinal effectiveness based on multi-wave test data is not provided by this analysis.

Table 1 provides summary descriptions of the evaluations that were included in the synthesis. The studies are listed by publication year, which roughly corresponds to the years of data collection. The table reveals that the early studies reported test scores in grade equivalents, employed fall-spring testing cycles, and assessed reading achievement only. The post-TIERS studies reported NCE gains in both reading and math and, generally, were based on annual testing cycles. Figure 1, which plots the frequency of observations by year of pretest, highlights the more abundant and consistent reports of test data since the implementation of TIERS in 1979.

Table 1 also displays the type of comparison group that was used in each study. As evident, the majority of studies (14 of 17 studies and 96.5 percent of the observations) relied on the norm-referenced model to evaluate the effectiveness of the program. Three of the studies utilized a control group that consisted of students not participating in the program. In the Glass, *et al.* (1970) report, these non-participants were enrolled in the same schools as the Title I students, but were either not considered in need of, or were not offered, programmatic services. The *Prospects* study reported by Puma, Jones, Rock, and Fernandez (1993) defined non-participants as those students enrolled in either non-eligible or eligible Title I schools who were not compensatory education participants. The *SES* made an attempt to identify non-participating students who were most similar to Title I students in respect to educational need. The researchers selected the comparison students by asking teachers at non-Title I schools to name students who were in most need of compensatory education had it been available at their schools. Although the study has been criticized on the basis that this comparison group may not have been truly equivalent (Slavin, *et al.*, 1989), follow-up analyses by Myers (1986), which corrected for selection bias, did not reveal results that substantiated these criticisms.

Table 1: Summary Descriptions of Studies (Listed in Order of Publication Date)

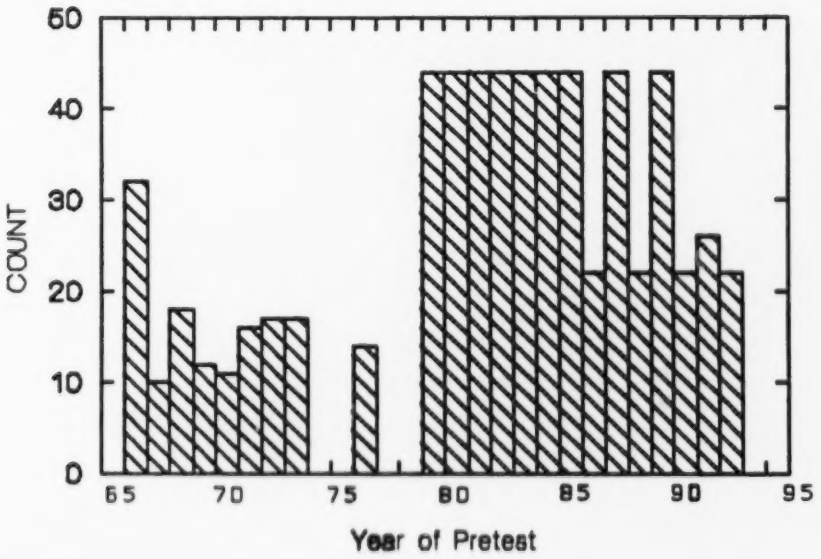
Primary Author	Years of Data Collection	Total Number of Title I Students Tested	Number of Observations[1]	Grades	Subjects	Original Test Score Metric	Testing Cycle	Comparison Type
DHEW (1967)	1966-67	20,990	32	2-7	Reading	Grade Equiv.	Fall-Spring	Norm-Referenced
Mosbaek (1968)	1967-68	1,145	10	3-12	Reading	Grade Equiv.	Fall-Spring	Norm-Referenced
Glass (1970)	1968-69	1,274	7	2,4,6	Reading	Grade Equiv.	Fall-Spring	All Non-Participants
Thomas (1976)	1968-73	2,697,567	70	1-12	Reading	Grade Equiv.	Fall-Spring	Norm-Referenced
Pelavin (1977)	1971-73	2,627	14	3-7	Reading	Grade Equiv.	Annual	Norm-Referenced
Frechtling (1978)	1976-77	2,761	2	1,3	Reading	Grade Equiv.	Fall-Spring	Norm-Referenced
Anderson (1982)	1979-80	3,717,742	44	2-12	Reading, Math	NCE	Fall-Spring, Annual	Norm-Referenced
Kennedy (1986)	1976-77	24,420	12	1-6	Reading, Math	NCE	Fall-Spring	Similar Needy
Gutmann (1988)	1980-86	17,322,096	264	2-12	Reading, Math	NCE	Fall-Spring, Annual	Norm-Referenced
Steele (1989)	1986-87	1,636,775	22	2-12	Reading, Math	NCE	Annual	Norm-Referenced
Sinclair (1990)	1987-88	2,683,217	44	2-12	Reading, Math	NCE	Fall-Spring, Annual	Norm-Referenced
Sinclair (1991)	1988-89	2,124,955	22	2-12	Reading, Math	NCE	Annual	Norm-Referenced
Sinclair (1992)	1989-90	4,224,603	44	2-12	Reading, Math	NCE	Fall-Spring, Annual	Norm-Referenced
Puma (1993)	1991-92	2,032	4	3,7	Reading, Math	NCE	Annual	All Non-Participants
Sinclair (1993)	1990-91	2,181,772	22	2-12	Reading, Math	NCE	Annual	Norm-Referenced
Sinclair (1994a)	1991-92	3,079,988	22	2-12	Reading, Math	NCE	Annual	Norm-Referenced
Sinclair (1994b)	1992-93	1,982,232	22	2-12	Reading, Math	NCE	Annual	Norm-Referenced
Totals:		41,706,196	657					

[1]Each observation corresponds to a particular comparison based on grade, subject, testing cycle, and year of pretest, from which an individual effect size was computed [e.g., 12 effect sizes were computed for the Kennedy, *et al.* (1986) study].

BEST COPY AVAILABLE

Furthermore. the study made the most extensive attempt to date at developing similar controls from a nationally representative sample.

Figure 1: Frequency of Observations Plotted by Year of Pretest



Results

Effect sizes were computed for observations using separate algorithms based on the comparison type and grade level. For norm-referenced comparison type observations, no test scores existed for control students. Instead, testing norms were employed as references. All TIERS data were reported in the NCE metric, and prior to TIERS all norm-referenced data were reported in the grade equivalent metric. Prior to the computation of effect sizes for norm-referenced comparisons, all data were transformed to the NCE metric. Only one of the studies that used grade equivalents, Pelavin

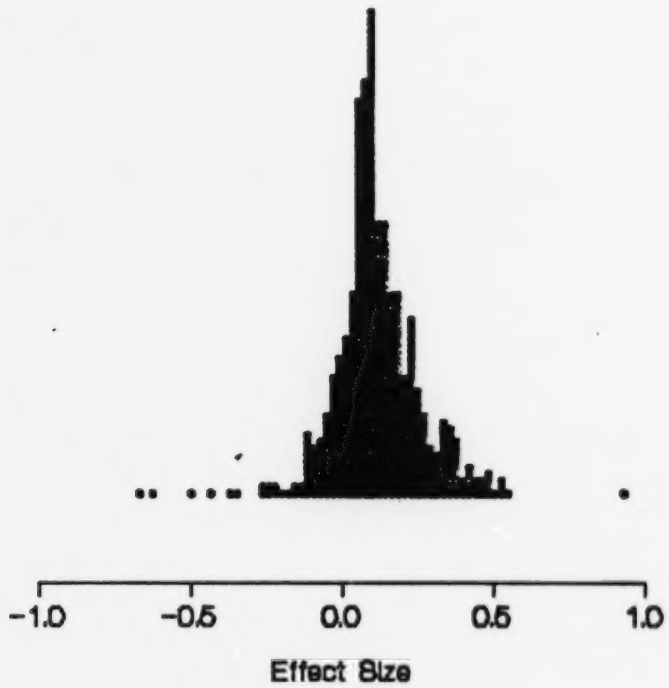
and David (1977), cited standard deviations for the sample. The other studies that reported grade equivalent data did not provide standard deviations, and generally did not report the test or tests that were used. The largest data source, Thomas and Pelavin's (1976) study, synthesized results from a number of different standardized tests. Because of these complications in defining the appropriate standard deviations to utilize, it was necessary to compute estimated standard deviations across all grade levels. A sample of reading and vocabulary tests that have been used in previous Title I evaluations was selected. The norming data for the 1964, 1984, 1985, and 1988 Iowa Test of Basic Skills (Lindquist and Hieronymus, 1964; Hieronymus and Hoover, 1990), the 1973 and 1982 Stanford Achievement Tests (Madden, Gardner, Rudman, Karlsen, and Merwin, 1973; Gardner, Rudman, Karlsen, and Merwin, 1982), the 1978 Gates-MacGinitie Reading Test (MacGinitie, Kamons, Kowalski, MacGinitie, and MacKay, 1978), and the 1976 California Achievement Test (CTB/McGraw-Hill, 1979) were consulted to ascertain standard deviation estimates. These standard deviations varied slightly across tests, but appreciably across grade level. Based on the data provided by the tests' norms, standard deviation estimates were derived by grade level utilizing a quadratic regression model. Finally, the standard deviation estimates were used in converting the grade equivalent data into the NCE metric.

After converting the grade equivalent data into the NCE metric, effect sizes for all norm-referenced comparisons were computed with an adjustment for artifacts due to regression that was based on the findings of Linn (1980) and Tallmadge (1982). As previously stated in the introduction, Linn (1980) and Tallmadge (1982) estimated the impact of regression on nationally-aggregated gains to be about 2 NCEs at second grade, and 1 NCE for third grade and above. Although some researchers [e.g., Linn, *et al.* (1982); Kennedy, *et al.*, (1986)] have claimed that other potential measurement artifacts and practical issues may bias norm-referenced reports of Title I students' achievement growth, there is no widely accepted estimation of their extent and magnitude.

Other formulae for effect sizes were employed for the few studies that included test data for control students. Separate formulae were specified for control group comparisons reported in the NCE metric and for those in the grade equivalent metric.

Figure 2 shows the distribution of the unweighted effect sizes for all 657 observations from the 17 studies. Clearly, the majority of observations were greater than 0, but most were modest in magnitude. One outlying effect size had a positive value of approximately 1.0, and negative outliers ranged from about -0.35 to -0.7.

Figure 2: Distribution of Effect Sizes



Preliminary analyses of the effect sizes were performed according to the assumptions of fixed effects models. Variances of the effect sizes for control group comparisons were computed based on the following formula:

$$v_i = ((N^T+N^C)/(N^T*N^C))+((\text{EFFECT SIZE}^2)/(2*(N^T+N^C)))$$

Given that there were no control students in the norm-referenced comparisons, the variance formula was:

$$v_i = ((1/N^T)+((\text{EFFECT SIZE}^2)/(2*(N^T)))$$

Unweighted mean effect sizes and overall weighted *T* values were computed for each of the 657 observations, for each comparison type (i.e., norm-referenced, similar needy students, and all non-participants), and for both testing cycles (i.e., fall-spring, and annual). Tables 2 and 3 present the results of these analyses.

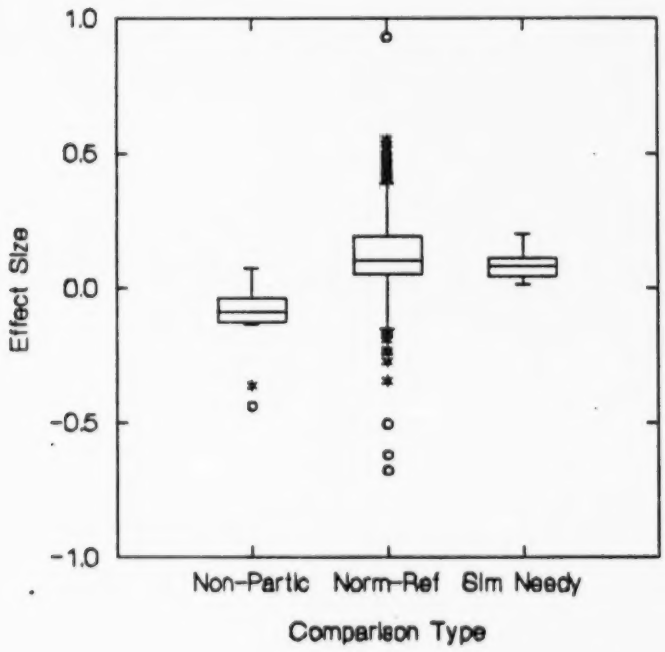
Table 2: Effect Sizes by Comparison Type

Comparison Type	Number of Observations	Mean Unweighted Effect Size (*d*)	Mean Weighted Effect Size (*T*)
Norm-Referenced Comparisons	634	0.121	0.167
Similar Needy Students	12	0.083	0.086
All Non-Participants	11	-0.121	-0.037
TOTAL	657	0.116	0.164

The results from the norm-referenced comparisons suggested that, even while adjusting for regression artifacts in NCE gains, the other controlled comparisons yielded relatively smaller mean effect sizes. Each mean *T* value computed in the tables was more positive than its corresponding unweighted mean effect size, indicating that more reliable estimates were generally more favorable estimates of the program's effect. The results for the norm-referenced comparisons suggested that Title I participants have achieved at an annual rate that exceeds the equipercentile expectation. In addition, participating students have posted larger annual gains than their similarly disadvantaged counterparts who did not receive Title I services. Nonetheless, Title I students have not gained at the same rate of their more advantaged peers.

Figure 3 provides a graphic illustration of the differences in the estimates of the 657 unweighted effect sizes that were produced by the three comparison types. The box and whisker plot reveals that there was more variability in the effect sizes from the norm-referenced comparisons. This may be due to the greater representation of both annual and fall-spring testing cycles within this comparison type, which led to differential estimates of the program's impact.

Figure 3: Box and Whisker Plot of Effect Size by Comparison Type



As evident from Table 3, the mean weighted effect size, *T*, for the fall-spring testing cycle (0.26) was close to three times larger than the mean *T* value for the annual testing cycle (0.10). Furthermore, as illustrated by the box and whisker plot of effect size by test cycle (see Figure 4), there was less variability in effect sizes based on annual testing cycles, which suggested that the summer months may have had a similar deleterious effect on Title I participants' achievement growth. Alternatively, greater variation in testing conditions may be associated with evaluations conducted on a

Table 3: Effect Sizes by Testing Cycle

Comparison Type	Number of Observations	Mean Unweighted Effect Size (*d*)	Mean Weighted Effect Size (*T*)
Annual Testing Cycle	348	0.069	0.103
Fall-to-Spring Cycle	309	0.170	0.262
TOTAL	657	0.116	0.164

fall-spring schedule. Most likely, both possible explanations contributed to these differences. As expected, effect sizes for the fall-spring data were greater than those for the minimal "sustained effect" measure provided by annual testing cycles.

Figure 4: Box and Whisker Plot of Effect Size by Testing Cycle

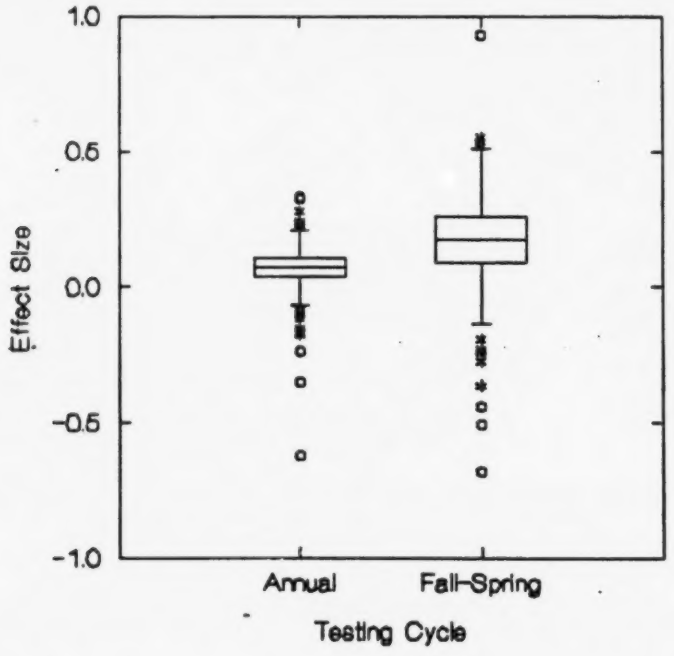
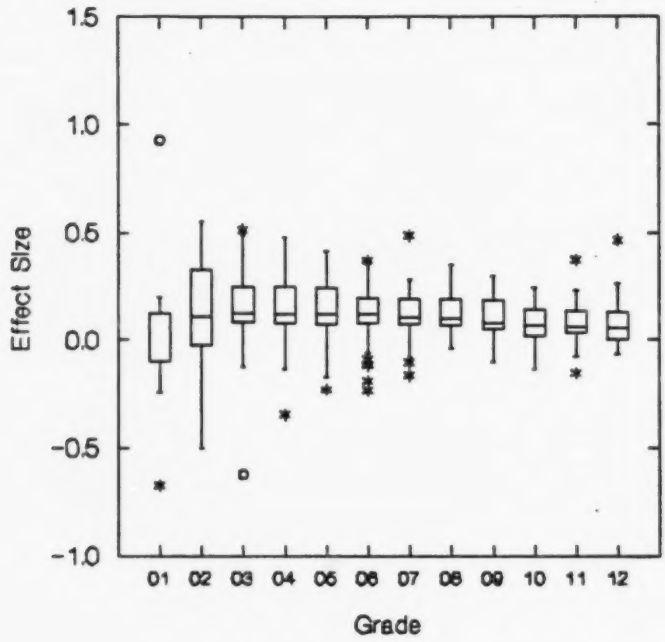


Figure 5 is a box and whisker plot of the effect sizes by grade level. As expected, this figure appears to show a moderate and relatively linear negative relationship between effect size and grade.

Most of the variability was across grades 2 through 7. In grades 8 through 12, the extent of variability within grades seemed to diminish. Both the negative relationship and the differences in the variability

Figure 5: Box and Whisker Plot of Effect Size by Grade

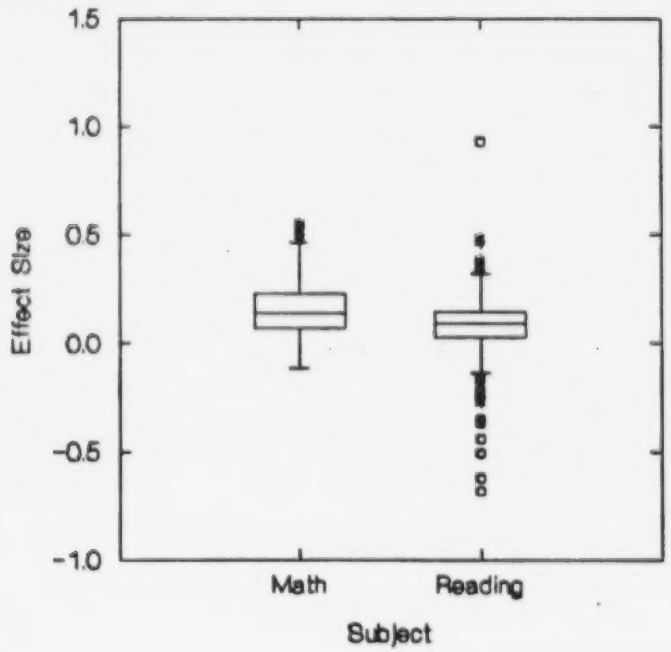


of the effect sizes across the grades may be a consequence of Title I selection practices. Because intermediate and high schools receive fewer funds than elementary schools, a relatively homogeneous group of the poorest performing students tend to receive services. Therefore, schools serving participants in the upper grades may experience more difficulty in devising programs to improve participants' academic standing. The greater degree of variability in the lower grades also may be due to the wider assortment of program options available in these grades. In addition, grades 2 through 7 were represented in a larger number of observations from a more diverse set of studies across a broader range of years. Few studies provided grade 1 observations (3 of the 17 studies), and no data were available for grade 1 students after the 1976-77 academic year. Although a greater number of

studies reported results for grades 8 through 12 (n = 11), most data for middle and high school participants came from evaluations completed after 1980. At least one of the grades from 2 through 7 was represented in all 17 studies (see Table 1).

The effect sizes for math and reading are plotted in Figure 6. As predicted, the math effect sizes were slightly larger than the reading effect sizes. This finding corroborated past research that has compared participants' outcomes by subject (Kennedy, *et al.*, 1986). The greater number of outlying effect sizes for reading may be related to the larger number of evaluations of Title I reading programs from a broader range of years. As indicated in Table 1, no evaluations prior to 1979-80 summarized Title I students' math achievement.

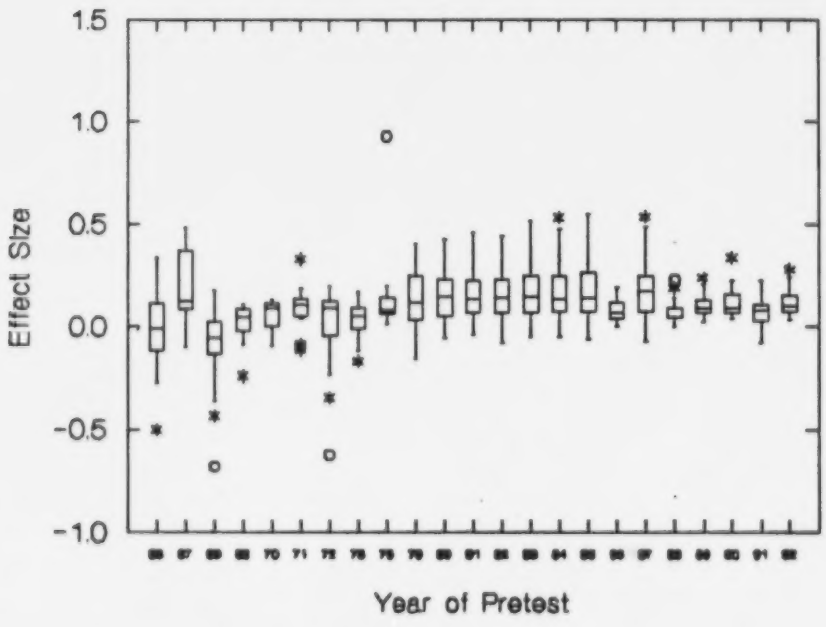
Figure 6: Box and Whisker Plot of Effect Size by Subject



The effect sizes are plotted by year of pretest in Figure 7. This figure generally seems to highlight a positive linear trend of the median unweighted effect sizes over the years. The clear

majority of the large positive outliers were from 1984-1992, and most outlying negative values were from the 1966-1973 school years. Except for the late 1960s, early 1970s, and 1991-92, each year was represented by a single study. Therefore, in all other years, the idiosyncracies of each study were at least partly responsible for the estimated effectiveness of the program within the particular year. Most of the observations during the period which appeared most stable, 1979 through 1992,

Figure 7: Box and Whisker Plot of Effect Size by Year of Pretest



were drawn from the state summaries of TIERS data, which used similar methods from year to year. Instability in the effect sizes across other years seemed to reflect differences in evaluation designs. For instance, as is evident from Table 1, evaluations of Title I during the early years focused on reading

programs exclusively. Because achievement gains for reading tend to be smaller than those for math, the smaller unweighted effect sizes from 1966 to 1976 may have been due to this methodological detail. Also, the studies conducted during the late 1980s and early 1990s employed annual testing cycles as an exclusive standard. This difference may explain the apparent smaller unweighted effect sizes from the 1990s. However, the potential positive impact of the exclusive use of fall-to-spring cycles during the first decade is not apparent in Figure 7.

As implied above, some of the box and whisker plots may be deceptive due to potential confounds among intermediary variables. This multiplicity of potential moderators underlies the concept of a study's true effect size as random (Raudenbush, 1994). To test whether the true effect size varied, in addition to the variability introduced by sampling variance, or estimation variance, a homogeneity test of the weighted effect size estimates was performed. Because, the value of 542,169.42 for the homogeneity test statistic, Q, exceeded the upper-tail critical value of chi-square at 656 degrees of freedom, the observed variance in the study effect sizes was significantly greater than that which would be expected by chance if all studies shared the same population effect size.

Therefore, in order to explain the variation in the study effect sizes, a weighted least squares multiple regression analysis for random effects was performed using effect size as the dependent measure and the intermediary variables as predictors. An estimate of the residual variance component was computed as the random effects variance plus the estimation variance, and weights were defined by the reciprocal of the residual variance component. Key interaction terms were included in the model as well. Table 4 presents the results of this analysis.

Calculations of the significance tests for the regression coefficients employed methods outlined by Hedges (1994). The standard error for each coefficient estimate generated by a conventional statistical software package was divided by the square root of the residual mean square from the

BEST COPY AVAILABLE

Table 4: Results of Weighted Least Squares Multiple Regression Analysis

DEP VAR: EFFECT SIZE N: 657 MULTIPLE R: 0.899 SQUARED MULTIPLE R: 0.808
ADJUSTED SQUARED MULTIPLE R: .805 STANDARD ERROR OF ESTIMATE: 8.955

VARIABLE	COEFFICIENT	STD ERROR	STD COEF	Z	P(2 TAIL)
SUBJECT	0.201	0.00190	0.712	105.79	0.000
GRADE	-0.010	0.00022	-0.393	-45.46	0.000
YEARPRE	0.012	0.00011	5.259	109.09	0.000
TESTCYCLE	-0.225	0.00190	-0.911	-118.42	0.000
COMPGRP	-0.188	0.00235	-0.157	-80.00	0.000
SUBJECT*GRADE	-0.015	0.00022	-0.406	-68.18	0.000
CYCLE*SUBJECT	-0.063	0.00156	-0.180	-40.38	0.000
CYCLE*GRADE	0.006	0.00022	0.193	27.27	0.000
CONSTANT	-0.625	0.00503	0	-124.25	0.000

ANALYSIS OF VARIANCE

SOURCE	SUM-OF-SQUARES	DF	MEAN-SQUARE	F-RATIO	P
REGRESSION	218317.072	9	24257.452	302.517	0.000
RESIDUAL	51960.158	648	80.185		

Variables entered:

GRADE: Grade level
SUBJECT: Reading=0, Math=1
YEARPRE: Year of pretest
TESTCYCLE: Fall-Spring=0, Annual=1
COMPGRP: Norm-Referenced Comparison=0, Control Group Comparison=1
SUBJECT*GRADE: SUBJECT * GRADE Interaction
CYCLE*SUBJECT: CYCLE * SUBJECT Interaction
CYCLE*GRADE: CYCLE * GRADE Interaction

analysis of variance for the regression. Nonetheless, because all coefficients were highly significant before this correction was made, it had a minimal impact on the final results. All intermediary variables along with the interaction terms accounted for about 80 percent of the variance in the weighted effect sizes. All predictors and interaction terms in the model were statistically significant.

Descriptions of the variables entered into the regression are provided in Table 4. The COMPGRP variable contrasted those observations based on the norm-referenced model to observations that were based on control group comparisons (i.e., non-participants and similar needy students). As indicated by the negative coefficient, control group comparisons yielded less positive estimates of the program's effectiveness than the norm-referenced model after controlling for the other predictors.

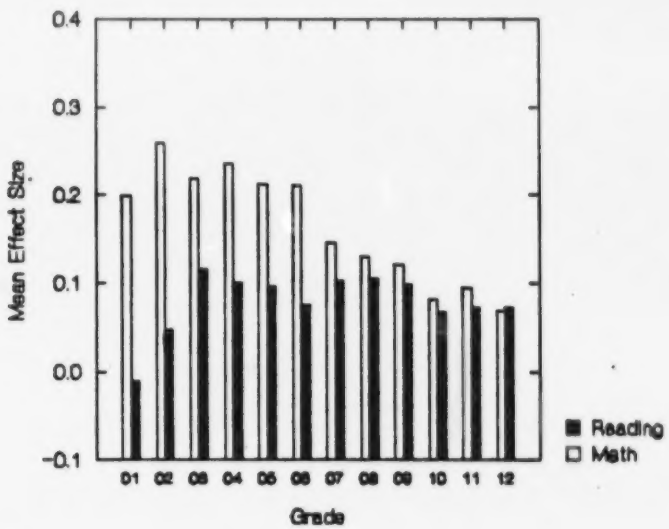
After controlling for the other variables, grade level was negatively related to effect size. Further, the negative coefficient for the test cycle type "dummy" variable suggested that the studies

BEST COPY AVAILABLE

based on annual gains produced smaller effect size estimates than those that used a fall-spring test cycle. The coefficient for the subject variable, which was positive, indicated that students participating in math programs tended to gain more than students in reading programs after holding the other variables constant.

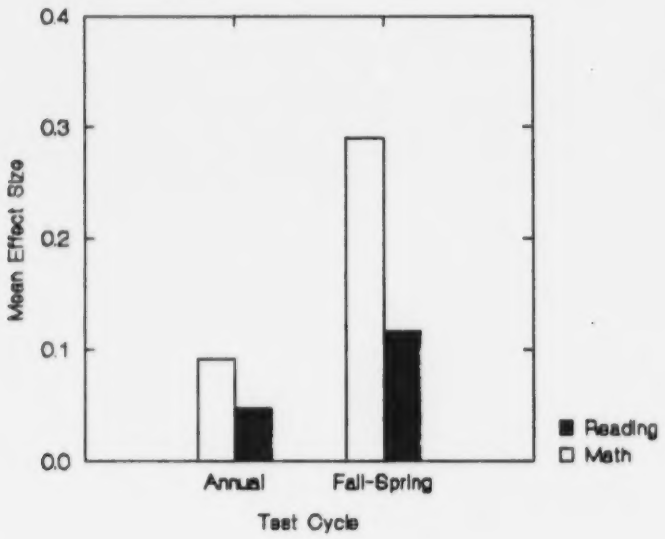
Finally, after controlling for the other variables, the program seemed to have become more effective as it matured. This finding was in contradiction to expectations and to widely held beliefs. However, no previous analyses have been designed to model the effect of Title I over the years of its operation while statistically holding constant several of the critical sources of variation in the evaluation approaches. Three interaction terms were included in the model: subject by grade, test cycle by subject, and test cycle by grade. Figure 8, which depicts the subject by grade interaction, shows that effect sizes for math were appreciably greater than those for reading in the elementary grades, but, this difference steadily diminished through the middle and high school grade levels.

Figure 8: Mean Effect Size as Function of Subject and Grade



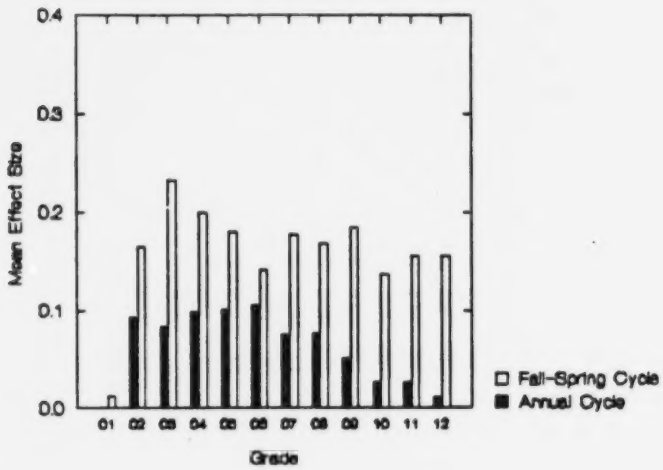
In Figure 9, which plots the subject by test cycle interaction, the math effect sizes were considerably greater than the reading effect sizes, but this difference decreased when the comparisons were based on annual testing cycles. This seemed to corroborate previous findings by Heyns (1977), which suggested that the summer effect was more deleterious to math than reading achievement.

Figure 9: Mean Effect Size as a Function of Subject and Testing Cycle



Finally, Figure 10 reveals that differences between the effect size estimates based on annual and fall-spring testing cycles were most pronounced in the early elementary grades and in the upper grades. This result suggested that the summer effect was most detrimental to disadvantaged students in early elementary and high schools. Interestingly, the widely held belief that Title I has a more profound impact in the early rather than later grades appeared to be supported by the distribution of effect sizes across grades for observations based on fall-spring testing cycles. However, as stated above, the effect sizes by grade level for the annual cycles indicated a different relationship.

Figure 10: Mean Effect Size as a Function of Grade and Testing Cycle



Conclusion

Title I is primarily a funding mechanism that allows for extensive variation, both across and within schools, in program design and implementation. Consequently, there is no single Title I "treatment." Although one purpose of this meta-analysis was to ascertain whether program services, taken as a whole, have had a significant impact on student achievement, it also addressed other questions about the program and the way in which it is evaluated. This analysis offered the possibility to examine if choices of different testing cycles and comparison methods have provided differential estimates of the program's impact. Also, this synthesis permitted investigation of the historical variation in the effect of Title I.

Much of the conventional wisdom regarding the program's effectiveness was supported by the findings, but there were some unexpected results. Due to the considerable heterogeneity in the effect size estimates, the mean weighted effect size of .164 should be interpreted with caution. A

BEST COPY AVAILABLE

combination of design and methodological differences in Title I programs and in the evaluation literature compromise the validity of generalizations regarding the overall effectiveness of Title I.

Not surprisingly, the program appeared to be effective for students who participated in the early elementary grades, but its impact diminished as the grade level increased. However, the interaction of grade level and testing cycle influenced the interpretation of this main effect. Further, Title I implementations seemed to be more beneficial in math than in reading, although the difference between programs in these subjects was not as pronounced in the upper grades. Those observations that employed fall-spring testing cycles tended to yield more favorable effect size estimates than those that used annual cycles. Also, evaluations that relied on norm-referenced comparisons were more likely to produce larger, more positive effect size estimates than studies that adopted non-equivalent control group designs.

Contrary to widely held beliefs regarding the historical stability of programmatic impact, the results suggested a positive trend for the educational effectiveness of Title I across the years of its operation. A number of factors may have contributed to this result. First, district and school staff have become more aware of the educational needs of at-risk students and, consequently, have become more committed to allocating resources to develop more constructive programs. Although many local educators and policymakers have developed this commitment toward improving the program independently, more stringent federal fiscal and procedural standards have had much to do with this evolution (Peterson, Rabe, and Wong, 1988). Second, besides greater concern for program compliance, federal policy has become more focused on developing strategies to encourage schools to improve their programs. This was reflected most notably by the 1988 Hawkins-Stafford Amendments, which offered schools greater flexibility in programmatic development in return for greater accountability for improved student outcomes. It was not possible to isolate the impact of each of

these factors on greater student achievement gains, but these findings appeared to indicate that some combination of them have contributed to this trend.

Although other artifacts may be related to the differences between annual and fall-spring gains, the interaction between testing cycle and grade seemed to reveal that the summer effect has been more deleterious to disadvantaged students in the early elementary grades and to adolescent students in the intermediate and upper grades. For the younger students, this may indicate the importance of the home environment during their primary developmental period. The distractions of early adulthood, which are prevalent in economically depressed communities, may be related to the substantially smaller annual gains of students participating in intermediate and upper grade programs.

These findings may point to the importance of allocating increased Title I dollars to schools in order to develop more extensive and effective summer programs targeted toward these students. However, for the older students, any supplemental educational intervention alone may not be adequate in counteracting the cumulative negative effects of their total educational environments. Finally, as suggested by the interaction of subject and testing cycle, which indicated that student gains in math may be more prone to the negative consequences of the summer effect, summer programs in math may be more important to participants' overall academic development than reading programs.

The evidence obtained from this analysis indicated that Title I has not fulfilled its original expectation: to close the achievement gap between at-risk students and their more advantaged peers. However, Title I alone cannot be expected to serve as the great equalizer. The results did suggest, however, that without the program it is likely that children served over the last 30 years would have fallen farther behind academically. If evaluated from this perspective, Title I has been an invaluable supplement to schools serving disadvantaged children.

References

Note: The 17 references included in the synthesis are marked by asterisks (*).

*Anderson, J.I. and Stonehill, R.M. (1982, March). *Measuring the effects of compensatory education--searching for convergence from national, state, and local evaluations.* Paper presented at the Annual Meeting of the American Educational Research Association, New York, NY.

Carter, L.F. (1984). The Sustaining Effects Study of compensatory and elementary education. *Educational Researcher, 13*(7), 4-13.

CTB/McGraw-Hill (1979). *California Achievement Tests, Technical Bulletin 1.* Monterey, CA: Author.

David, J.L. and Pelavin, S.H. (1977). *Research on the effectiveness of compensatory education programs: A reanalysis of data.* Menlo Park, CA: SRI International.

Davis, A. (1991). Upping the stakes: Using gain scores to judge program effectiveness in Chapter 1. *Educational Evaluation and Policy Analysis. 13,* 380-388.

Trismen, D.A., Waller, M.I., and Wilder, G. (1975). *A descriptive and analytic study of compensatory reading programs.* Princeton, N.J.: Educational Testing Service.

Farrar, E. and Millsap, M.A. (1986). *State and local administration of the Chapter 1 program, Volume 1 and Appendix Volume: Issue summaries and methodology.* Cambridge, MA: Abt Associates, Inc.

Forbes, R.H. (1985). Academic achievement of historically lower-achieving students during the seventies. *Phi Delta Kappan, 66,* 542-544.

*Frechtling, J., *et al.* (1977). *The effects of services on student development.* Washington, DC: National Institute of Education.

Gardner, E.F., Rudman, H.C., Karlsen, B., Merwin, J.C. (1982). *Stanford Achievement Test: Scoring manual.* New York, NY: Harcourt Brace.

*Glass, G.V. (1970). *Data analysis of the 1968-69 survey of compensatory education (Title I).* Washington, D.C.: U.S. Department of Health, Education, And Welfare.

*Gutmann, B. and Henderson, A. (1988). *A summary of state Chapter 1 participation and achievement information for 1985-86.* Washington, DC: Decisions Resources Corporation.

Hayes, D.P. and Grether, J. (1969, April). *The school year and vacations: When do students learn?* Revision of a paper presented at the Eastern Sociological Association Convention, New York, NY.

Hedges, L.V. (1994). Fixed effects models. In H. Cooper and L.V. Hedges (Eds.). *The handbook of research synthesis.* (pp. 285-289). New York, NY: Russell Sage Foundation.

Hieronymus, A.N., and Hoover, H.D. (1990). *Iowa Test of Basic Skills, Manual for School Administrators*. Chicago, IL: Riverside Publishing Company.

*Pelavin, S.H., and David, J.L. (1977). *Evaluating long-term achievement: An analysis of longitudinal data from compensatory education programs*. Menlo Park, CA: Stanford Research Institute.

Gordon, E.W. (1970). Compensatory education: Evaluation in perspective. *Information Retrieval Center on the Disadvantaged Bulletin*, Volume VI, No. 5.

Heyns, B. (1978). *Summer learning and the effects of schooling*. New York, NY: Academic Press.

Jaeger, R.M. (1978). The effect of test selection on Title I project impact. *Educational Evaluation and Policy Analysis, 1*, 33-40.

Kaskowitz, D.H. and Norwood, C.R. (1977). *A study of the norm-referenced procedure for evaluating project effectiveness as applied in the evaluation of project information packages*. Menlo Park, CA: Stanford Research Institute.

*Kennedy, M.M., *et al.* (1986). *The effectiveness of Chapter 1 services. Second interim report from the national assessment of Chapter 1*. Washington, D.C.: Office of Educational Research and Improvement.

Lindquist, E.F. and Hieronymus, A.N. (1964). *The Iowa Test of Basic Skills: Manual for administrators, supervisors, and counselors. Multi-level edition for grades 3-9*. Chicago, IL: Riverside Publishing Company.

Linn, R.L. (1979). Validity of inferences based on the proposed Title I evaluation models. *Educational Evaluation and Policy Analysis, 1*, 23-32.

Linn, R.L. (1980). Regression towards the mean and the interval between test administrations. *New Directions for Testing and Measurement, 8*, 59-82.

Linn, R.L., Dunbar, S.B., Harnisch, D.L., and Hastings, C.N. (1982). The validity of the Title I evaluation and reporting system. In E. House, S. Mathison, J. Pearsol, and H. Preskill (Eds.). *Evaluation studies review annual*. Beverly Hills, CA: Sage Publications.

MacGinitie, W.H., Kamons, J., Kowalski, R.L., MacGinitie, R.K., and MacKay, T. (1978). *Gates-MacGinitie reading tests, teacher's manual*. Boston, MA: Houghton Mifflin.

Madden, R., Gardner, E., Rudman, H., Karlsen, B., and Merwin, J. (1973). *Stanford achievement test norm booklet, form A, primary level I to intermediate level I battery*. New York, NY: Harcourt Brace Jovanovich, Inc.

Martin, R., and McClure, P. (1969). *Title I of ESEA: Is it helping poor children?* Washington, DC: Washington Research Project and NAACP Legal Defense and Educational Fund, Inc.

McLaughlin, D.H. (1977). *Title I, 1965-1975: Synthesis of the findings of federal studies*. Palo Alto, CA: American Institute for Research.

*Mosbaek, E.J. (1968). *Analysis of compensatory education in five school districts, volume I: Summary. Final report.* Washington, DC: General Electric Company--TEMPO.

Mullin, S.P. and Summers, A.A. (1983). Is more better? The effectiveness of spending on compensatory education. *Phi Delta Kappan, 64*(5), 339-347.

Myers, D.E. (1986). *An analysis of the impact of Title I on reading and math achievement of elementary school aged children. Revised.* Washington, DC: Decisions Resources Corporation.

*Office of Education, Department of Health, Education, and Welfare (1967). *Second annual report of Title I of the Elementary and Secondary Education Act of 1965 school year 1966-67.* Washington, DC: Author.

Peterson, P.E., Rabe, B.G., and Wong, K.W. (1986). The evolution of the compensatory education program. In Doyle and Cooper, (Eds.). *Federal aid to the disadvantaged.*

*Puma, M.J., Jones, C.C., Rock, D., and Fernandez, R. (1993). *Prospects: The congressionally mandated study of educational growth and opportunity, interim report.* Bethesda, MD: Abt Associates.

Raudenbush, S.W. (1994). Random effects models. In H. Cooper and L.V. Hedges (Eds.). *The handbook of research synthesis.* (pp. 301-321). New York, NY: Russell Sage Foundation.

Rossi, R.J. and McLaughlin, D.H., Campbell, E.A., and Everett, B.E. (1977). *Summaries of major Title I evaluations, 1966-1976.* Palo Alto, CA: American Institute for Research.

*Sinclair, B., and Gutmann, B. (1990). *A summary of state Chapter 1 participation and achievement information -- 1987-88.* Rockville, MD: Westat, Inc.

*Sinclair, B. and Gutmann, B. (1991). *A summary of state Chapter 1 participation and achievement information -- 1988-89.* Rockville, MD: Westat, Inc.

*Sinclair, B., and Gutmann, B. (1992). *A summary of state Chapter 1 participation and achievement information -- 1989-90.* Rockville, MD: Westat, Inc.

*Sinclair, B., and Gutmann, B. (1993). *A summary of state Chapter 1 participation and achievement information -- 1990-91.* Rockville, MD: Westat, Inc.

*Sinclair, B. and Gutmann, B. (1994a). *State Chapter 1 participation and achievement information -- 1991-92.* Rockville, MD: Westat, Inc.

*Sinclair, B., and Gutmann, B. (1994b). *State Chapter 1 participation and achievement information -- 1992-93.* Rockville, MD: Westat, Inc.

Slavin, R.E., Karweit, N.L., and Madden, N.A. (1989). *Effective programs for students at-risk.* Boston, MA: Allyn and Bacon.

*Steele, D. and Gutmann, B. (1989). *A summary of state Chapter 1 participation and achievement*

information for 1986-87. Washington, DC: Decisions Resources Corporation.

Tallmadge, G.K. (1982). An empirical assessment of norm-referenced evaluation methodology. *Journal of Educational Measurement, 19,* 97-112.

Tallmadge, G.K. and Wood, C.T. (1981). *User's guide to the ESEA Title I evaluation and reporting system.* Mountain View, CA: RMC Research Corporation.

*Thomas, T.C., and Pelavin, S.H. (1976). *Patterns in ESEA Title I reading achievement.* Menlo Park, CA: Stanford Research Institute.

Timpane, M. (1976). Evaluating Title I again? In C. Abt (Ed.). *The evaluation of social programs,* (pp. 415-423). Beverly Hills, CA: Sage.

Wargo, M.J., Tallmadge, G.K., Michaels, D.D., Lipe, D., and Morris, S.J. (1972). *ESEA Title I: A reanalysis and synthesis of evaluation data from fiscal year 1965 through 1970.* Palo Alto, CA: American Institute for Research.